

PATENT APPLICATION

MULTI-STAGE SWITCHING FOR NETWORKS

Inventor: **DOUGLAS J. VANESKO**
6807 Meadow Road
Dallas, Texas 97230
a Citizen of the United States of America

Assignee: **CISCO TECHNOLOGY, INC.**
170 West Tasman Drive
San Jose, California 95134
a Corporation of the state of California

Status: Large Entity

RITTER, LANG & KAPLAN LLP
12930 Saratoga Ave., Suite D1
Saratoga, CA 95070
Tel: 408-446-8690

PATENT

5

MULTI-STAGE SWITCHING FOR NETWORKS

This application claims priority to provisional Application No. 60/299,714, filed June 18, 2001, which is hereby incorporated by reference.

10

BACKGROUND OF THE INVENTION

The present invention relates to multi-stage switching. More specifically, the invention relates to multi-stage grooming switches for use in networks.

15

“Switching” refers to the process of transmitting data from a set of input ports to a set of output ports. With multi-stage switching, the switching fabric includes multiple stages or nodes between the input ports and the output ports. Each node can be implemented in a variety of ways including a full N^2 cross-connect or a $2N$ Clos network.

A Clos network or matrix is a multi-stage interconnect where each switch in each stage is connected to each switch in the successive stage by exactly one link. For example, an asymmetric $2N$ Clos network has N inputs and outputs and $2N$ center states. In general,

“blocking” occurs when a connection requirement is not fulfilled. With Clos networks, there are two types of blocking: rearrangeable non-blocking and non-rearrangeable non-blocking.

Rearrangeable non-blocking refers to blocking where the existing connections can be moved to different center stages such that a center stage switch becomes available to make the desired

5 connection. Non-rearrangeable blocking refers to blocking where a connection that is blocked cannot be created with the current set of existing connections. A $2N$ Clos network has been proven to be rearrangeably non-blocking for bi-cast (i.e., meaning that a signal can be output on two different links or outputs). When a signal is tri-cast, there is a possibility that the Clos network will be non-rearrangeably blocked. The probability of this happening can be called the
10 non-rearrangeable blocking probability.

As communication networks continue to grow, there is an increasing need to design multi-stage switches that can handle significant increases in bandwidth and efficient to manufacture and operate. Additionally, it would be beneficial to provide a method of making connections through the switching fabric and reduces or eliminates the probability of

15 blocking.

SUMMARY OF THE INVENTION

The present invention provides innovative techniques for multi-stage switching in networks. In general, a multi-stage network element switch can include input nodes, center nodes and output nodes. Each node has inputs and outputs, with the signals passing from the input nodes to the center nodes to the output nodes and out. By adjusting the switching capabilities of each node, the network element switch is able to be configured as desired. Some specific embodiments of the invention are described below.

In one embodiment, the invention provides a network element switch. The network element switch includes a pair of input nodes, with each input node having $2n$ inputs and $2n$ outputs (n is ≥ 1). Additionally, the network element switch includes a pair of center nodes, with each center node receiving n inputs from each of the input nodes and having $2n$ outputs. The network element switch includes a pair of output nodes, with each output nodes receiving n inputs from each of the center nodes and having $2n$ outputs. Lastly, an any-to-any mapping is not guaranteed from the pair of input nodes to the pair of output nodes. In one embodiment, each node is a $2N$ Clos matrix.

In another embodiment, the invention provides a network element switch. The network element switch includes a pair of input nodes, with each input node having $2n$ inputs and $2n$ outputs (n is ≥ 1). Additionally, the network element switch includes a pair of center nodes, with each center node receiving n inputs from each of the input nodes and having $2n$ outputs. The network element switch includes a pair of output nodes, with each output nodes

receiving n inputs from each of the center nodes and having $2n$ outputs. Lastly, mapping links for signals between nodes is selected to minimize the differences between the number links between the nodes.

In another embodiment, the invention provides a method of mapping signals across a network element switch having input nodes, center nodes and output nodes. For each possible mapping of a specific signal across the network element switch, the number of signals between each of the input nodes and center nodes is calculated. Additionally, for each possible mapping of the specific signal across the network element switch, the number of signals between each of the center nodes and output nodes is calculated. The mapping of the specific signal across the network element switch is selected where the difference in the number of signals between each of the input nodes and center nodes is minimized and the difference in the number of signals between each of the center nodes and output nodes is minimized.

Other features and advantages of the invention will become readily apparent upon review of the following description in association with the accompanying drawings, where the same or similar structures are designated with the same reference numerals.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows an example of a network including multiple network elements and a network element switch.

FIG. 2 shows Clos matrix that is non-rearrangably blocked.

5 FIG. 3 shows a Clos matrix that is similar to FIG. 2 except that non-blocking tri-cast has been obtained for inputs 1-4.

FIG. 4 shows an example of an asymmetric $2N$ Clos matrix where $N=48$.

FIGS. 5A and 5B show six chip Clos matrices that can be utilized to implement the Clos matrix shown in FIG. 4.

10 FIGS. 6A-6C show examples of making connections through a six chip Clos matrix and FIG. 6D show a table that illustrates the affects of different configurations.

FIG. 7 shows a flowchart of a process of mapping signals across a network element switch having input nodes, center nodes and output nodes.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

In the description that follows, the present invention will be described in reference to embodiments that perform switching in network elements. However, embodiments of the invention are not limited to any particular environment, application or implementation. For example, six chip multi-stage embodiments will be described, but the invention may be advantageously applied to other configurations, including those that use fewer or more chips and/or stages. Therefore, the description of the embodiments that follows is for purposes of illustration and not limitation.

FIG. 1 shows an example of a network including a network element switch. The network shown is intended to illustrate exemplary network element switches that can utilize embodiments of the invention. However, the invention is not limited to the network shown and can be utilized in other network configurations.

An optical core 101 represents a high-speed optical network. Optical core 101 can utilize optical network elements such as the Cisco ONS 15454 available from Cisco Systems, Inc., San Jose, CA. In general, the optical core represents the next generation fiber optics network.

Within a metropolitan area, metro network elements 103 can transmit network data within a ring such as a Synchronized Optical Networks (SONET) ring. An example of a metro element 103 is the Cisco ONS 15454 available from Cisco Systems, Inc., San Jose, CA.

A cross-connect (XC) 105 provides switching and allows data to be transmitted between optical core 101 and metro elements 103. As a bandwidth of optical core 101 is typically higher than the metro element ring, cross-connect 105 can act to aggregate multiple data signals for use in the higher bandwidth network. Cross-connect 105 is just one example
5 of a network element switch that can utilize embodiments of the invention.

An edge device 107 transmits network data to and from customer premise equipment (CPE) 109. Edge device 107 is a network element that, among other things, receives network data for one of the customer premise equipment 109 and routes the network data to the appropriate destination. An example of edge device 107 is the Cisco ONS 15327 available
10 from Cisco Systems Inc., San Jose, CA. Although this network description is (or is expected to be) fairly common, embodiments of the invention are not limited to the network as has been described.

Embodiments of the invention utilize a multi-stage switching fabric. Each stage typically includes multiple nodes, where the nodes perform switching of the signals at that
15 stage. As mentioned previously, each node can be implemented as a full N^2 cross-connect or a $2N$ Clos matrix. Since some embodiments utilize $2N$ Clos matrices, it may be beneficial to discuss these initially.

FIG. 2 shows an example of a $2N$ Clos matrix exhibiting non-rearrangeable blocking. As shown, the $2N$ Clos matrix has $N = 4$. Thus, the nodes at the first stage include 4 inputs
20 (N), the center stage includes 8 nodes ($2*N$), and the third stage includes nodes that include 4

outputs (N). Although there is only one connection between nodes of different stages, for simplicity, not all the connections between the nodes are shown.

Assume that it is attempted to tri-cast inputs 1, 2 and 3. As shown, each of these signals is tri-cast to the second stage. However, utilizing this technique results in there not being an available center stage node for the last signal 3 (see the circled output). Non-rearrangeable blocking has occurred because a center stage node is not available to complete a connection. No matter how the other connections are rearranged, no common center stage node can be found.

With embodiments of the invention, the restriction of any input to any output (any-to-any mapping) is relaxed in order to guarantee non-blocking. In general, to obtain non-blocking multi-cast, the inputs to and outputs from the center stage are controlled.

FIG. 3 shows a Clos matrix that is similar to FIG. 2 except that non-blocking tri-cast has been obtained for inputs 1, 2, 3, and 4. As shown, each input signal has been restricted to use no more than two center nodes and non-blocking multicast has been achieved. Thus, by relaxing any-to-any mapping restriction, multicast is accomplished with no blocking probability.

FIG. 4 shows asymmetric 2N Clos network where $N = 48$. As shown, there are 48 inputs and outputs, and 96 center nodes ($2*N$). Additionally, there are 128 first and third stage nodes. The 2N Clos matrix shown in FIG. 4 can be implemented utilizing six chips (or grooming chips) connected as shown in FIGS. 5A and 5B. Each chip (identified as chips #0-

5) has $2N$ inputs and outputs. There are $2(N-1)$ connections between each chip of each stage. The six chip embodiments in FIGS. 5A and 5B are for $N = 6$. The resulting switching fabric allows for rearrangably non-blocking up to 64-cast on each output chip (total of 128 possible outputs on both output chips). The differences between FIGS. 5A and 5B will be discussed later.

Each chip can be either a full N^2 cross-connect or a $2N$ Clos matrix. For purposes of discussion, each chip will be assumed to be implemented in a single chip and be an asymmetric $2N$ Clos matrix. In a preferred embodiment, the six chips are identical and are the VC 2002 Grooming Switch available from Velio Communications, Milpitas, CA. Each Velio 2002 Chip is a 72×72 $2N$ Clos matrix. Multi-chip implementation of a grooming switch and full N^2 grooming cross-connects equally apply.

Using traditional analysis, the six chip Clos matrix shown in FIGS. 5A and 5B would be limited to a maximum of tri-cast and maintain a zero blocking probability. The first stage would be restricted to uni-cast, otherwise there would be a blocking probability exiting the first stage. The second and third stages could each perform bi-cast. A signal could be quad-cast if all four outputs were on a same output chip or if the outputs were split 2/2 among the chips. If however, the outputs were split 3/1, it would require tri-cast from either the second or third stage chips which would introduce a blocking probability.

As mentioned above, the multi-casting restrictions of the Clos network are a result of mapping any input to any output. If the mappings were controlled to avoid the blocking cases, a much higher degree of multi-casting can be achieved. With an embodiment of the

invention, the inputs to and outputs from the second stage are carefully controlled to avoid blocking. In this manner, non-blocking multi-cast of the number of single chip ports times 2 (ports*2) can be achieved. With respect to FIG. 5A, this translates to a multi-cast of 128 to each chip.

5 Now referring to FIG. 5A which shows a six chip Clos matrix, there are two groups of inputs to and two groups of outputs from each second stage chip. The input groups, uni-links, route signals from chip #0 and chip #1. The output groups, multi-links, route signals to chip #4 and chip #5. Preferably, both inputs and outputs on each second stage chip are traversed in a round-robin fashion to insure that center stage nodes are used efficiently. If a signal is
10 desired on X outputs of a third stage chip, it should appear $(X + 1)/2$ times on its input (i.e., restrict the third stage to bi-cast).

The first stage chips provide the appropriate inputs to the second stage chips. A given signal is sent to either chip #2 or chip #3, not both. The first stage is uni-cast, which is strictly non-blocking on a $2N$ Clos matrix.

15 If each input signal is restricted to use no more than two center stage nodes (see FIG. 4), non-blocking can be achieved. To obtain non-blocking multi-cast, the inputs to and outputs from the center stage are controlled. The center stage's job is to route signals to the appropriate third stage chip. When creating a circuit, the input and output chips are fixed. There are two different paths a signal can take. The signal may be routed through second
20 stage chip #2 or chip #3. The key is that the second stage chip only has to deliver the signal to the appropriate third stage chip. It is not required that the signal be routed on any particular

port/STS combination. The only restriction is that the signals are routed to the correct output chip.

The third stage is responsible for routing the signals received from the second stage chips to the desired output locations. This stage is also required to perform a bi-cast of signals
5 if required. A $2N$ Clos matrix is rearrangably non-blocking for bi-cast.

When the matrix fabric is updated, several options are available. Rearrangement of the circuits could be performed. However, this could be very complicated. Given the processing power available on many switching fabrics, a simpler solution may be to recalculate the entire fabric. An advantage of recalculating the matrix is that the time required
10 is likely to be constant regardless of the number of changes made.

A six chip switching network includes six cross-connected chips arranged in three stages. Two input chips constitute the first stage and each input chip includes $2N$ inputs for a total of $4N$ inputs where $N \geq 1$. Similarly, two output chips constitute the output stage with each output chip having $2N$ outputs totaling $4N$ outputs for the network. The center stage also
15 includes two chips with each chip connected to chips in the other stages by N links.

The difference between the networks shown in FIGS. 5A and 5B are that the first stage is uni-cast in FIG. 5A and the first stage is bi-cast in FIG. 5B. The network shown in FIG. 5A would not be able to support arbitrary multi-cast. However, because the network shown in FIG. 5B utilizes bi-cast out of the input stage, arbitrary multi-cast can be supported. It should
20 be noted that the six chip network is rearrangably non-blocking for multi-cast with the input

and output stages restricted to bi-cast and the center stage capable of multi-cast. Uni-cast from the input stage is sufficient if multi-cast is limited to $(2N-1)$.

Now that a six chip network has been described, it may be beneficial to describe a process for mapping signals across the network. FIG. 6A shows a six chip network with 2N inputs and 2N outputs with $N=2$. As shown, the signals are identified as signals 1-8. Additionally, the desired signals on the outputs are shown.

In order to obtain the desired outputs, signal 1 has been routed through chip #0 to chip #2 and then bi-cast from chip #2 to chips #4 and #5. Signal 5 has been routed through chip #1 through chip #3 and then bi-cast from chip #3 to chips #4 and #5.

Assume now that the next signal to be connected through the network shown in FIG. 6A is signal 2. FIGS. 6B and 6C show two different routing configurations for signal 2.

In FIG. 6B, signal 2 is routed through chip #0 to chip #2 and then to chip #4 and its output. In contrast, FIG. 6 shows that signal 2 has been routed through chip #0 to chip #3 and then to chip #4 and its output.

In one embodiment of the invention, the routing for a specific signal is chosen to minimize the differences between the number of links between the stages. In order to better illustrate this technique, FIG. 6D shows an example of differences in the number of links between the nodes shown in FIGS. 6B and 6C.

Referring to the table in FIG. 6D, the first column represents the difference in the number of links between the chips in the first stage and chips in the second stage. The second column shows differences in the number of links between the number of links between chips of the second stage and chips of the third stage.

5 The first row represents the differences for FIG. 6B where signal 2 is routed through the top center node. As shown in the table, the number of links between chip #0 and chip #2 is 2, the number of links between chip #0 and chip #3 is 0, the number of links between chip #1 and chip #2 is 0 and the number of links between chip #1 and chip #3 is 1. Since the greatest number of links is 2 and the least number of links is 0, the table shows that the
10 difference in links between the first stage chips and the second stage chips is 2. In a similar manner, the table shows that the difference in the number of links between the second stage chips and the third stage chips is 1 for FIG. 6B.

When the links between the chips in the different stages are counted and the difference is calculated as described above, FIG. 6D shows that the difference in the number of links
15 between the first stage chips and the second stage chips is 1 when the signal 2 is routed through the bottom center node. Also, the difference in the number of links between the second stage chips and the third stage chips when the signal is routed through the bottom center node is 1.

Accordingly, signal 2 would be routed through the bottom center node as shown in
20 FIG. 6C because the differences between the number of links between the chips of the different stages is minimized (e.g., the difference between the number of links between the

chips of the first and second stages instead of two). In general, this methodology attempts to utilize all available paths are evenly as possible.

Now that an example has been described, FIG. 7 shows a flow chart of a process of mapping signals across a network element switch having input nodes, center nodes and output nodes. At a step 501, the number of signals between each of the input nodes and center nodes is calculated. For each possible mapping of a specific signal across the network element switch, the number of signals between each of the input nodes and center nodes is calculated.

At a step 503, the number of signals between each of the center nodes and output nodes is calculated. As before, for each possible mapping of the specific signal across the grooming switch, the number of signals between each of the center nodes and output nodes is calculated.

The mapping across the network element switch is selected that minimizes the differences in the number of signals at a step 505. Thus, the mapping of the specific signal across the network element switch is selected where the difference in the number of signals between each of the input nodes and center nodes is minimized and the difference in the number of signals between each of the center nodes and output nodes is minimized.

Although in the preferred embodiments the switch fabric utilizes three stages, other embodiments can use fewer or more stages. Additionally, FIG. 7 shows an embodiment that minimizes the difference in the number of signals between both the first/center stages and the

center/last stage; however, other embodiments can minimize the difference in the number of signals between only two stages or more than two stages.

While the above is a complete description of preferred embodiments of the invention, various alternatives, modifications, and equivalents can be used. It should be evident that the invention is equally applicable by making appropriate modifications to the embodiments described above. Therefore, the above description should not be taken as limiting the scope of the invention that is defined by the metes and bounds of the appended claims along with their full scope of equivalents.